DOCUMENT RESUME

ED 178 562 TM 009 694

AUTHOR Roid, Gale; Haladyna, Tom
TITLE A Review of Item Writing Methods for Criterion-Referenced Tests in the Cognitive Domain.
PUB DATE Oct 78
NOTE 42p.; Paper presented at the Annual Meeting of the Military Testing Association (Oklahoma City, OK, October, 1978)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Behavioral Objectives; *Cognitive Objectives; *Cognitive Tests; Computer Programs; Concept Formation; Course Content; *Criterion Referenced Tests; Guidelines; Linguistic Patterns; *Test Construction; Test Items
IDENTIFIERS Computer Assisted Test Construction; *Domain Referenced Tests

ABSTRACT

Item writing methods for criterion referenced tests in the cognitive domain are reviewed. The methods are classified on a continuum from informal-subjective to automated-objective. Examples of approaches to improving instructional objectives are given. Domain-based tests are derived from content specifications and five approaches are described: (1) item forms; (2) linguistic-based approaches; (3) facet theory; (4) concept-based testing; and (5) computer-based methods. Rules which govern item construction and which specify the structure, format, and some of the wording are called item forms. Linguistic-based items are based on directions to the item writer on how to rearrange segments of the instruction to obtain a certain item type. Facet theory is viewed as a way to develop a population of items representing the domain of instruction. Concept-based testing is used to test a student's understanding of a concept by checking for generalization to new examples and discrimination of non-examples. Computer systems for creating domain-based criterion referenced tests are available. Areas for improvement are described as the definition of domains and creation of item-writing algorithms, both at the conceptual and factual levels. Bibliographic citations are appended. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED178562

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Gale Roid

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A REVIEW OF ITEM WRITING METHODS FOR CRITERION-REFERENCED TESTS IN THE COGNITIVE DOMAIN

Gale Roid and Tom Haladyna
Teaching Research Division
Oregon State System of Higher Education
Monmouth, Oregon 97361

Abstract

This paper provides a review of the emerging technology of test-item writing for criterion-referenced tests. Several different approaches to item development are discussed. A continuum of item writing methods is proposed ranging from informal-subjective methods to automated-objective methods. Examples of techniques include objective-based item writing, amplified objectives, item forms, facet design, domain-referenced concept testing and computerized techniques. Data from the authors' studies of item writing techniques are also reviewed. Recommendations for further research and for applications to military testing are presented.

TM009 694

A paper presented at the annual meeting of the Military Testing Association, Oklahoma City, October 1978.

A REVIEW OF ITEM WRITING METHODS FOR CRITERION-REFERENCED TESTS IN THE COGNITIVE DOMAIN

Gale Roid and Tom Haladyna
Teaching Research Division
Oregon State System of Higher Education
Monmouth, Oregon 97361

Since the seminal papers by Glaser (1963) and Popham and Husek (1969), there has been an astounding number of studies of criterion-referenced (CR) testing. In a review of technical issues relating to CR testing, Hambleton, Swaminathan, Algina, and Coulson (1978) indicate the existence of over 600 references, which is certainly a reflection of the importance that is placed on CR testing.

Although much has been written on the development of CR tests, it has only been recently that a well-defined methodology has emerged which can be generally described as a "theory" of CR testing. Millman (1974a) and Hambleton, et al. (1978) provide very comprehensive state-of-the-art reviews which capture the essence of CR testing.

Developing CR Tests

Five steps are identified that are essential aspects of CR test development (illustrated in Figure 1). The first step is the conceptualization of the content to be learned. Initially the instructional developer or instructor must identify what the student must learn as a consequence of instruction. This step may be somewhat introspective. It is a "private event" that is purely abstract, but it is a vital beginning in the process of planning instruction and the associated CR testing that is part of this instruction.

Following the conceptualization of what is to be taught, in step two instructional intent is then transformed into either (a) instructional

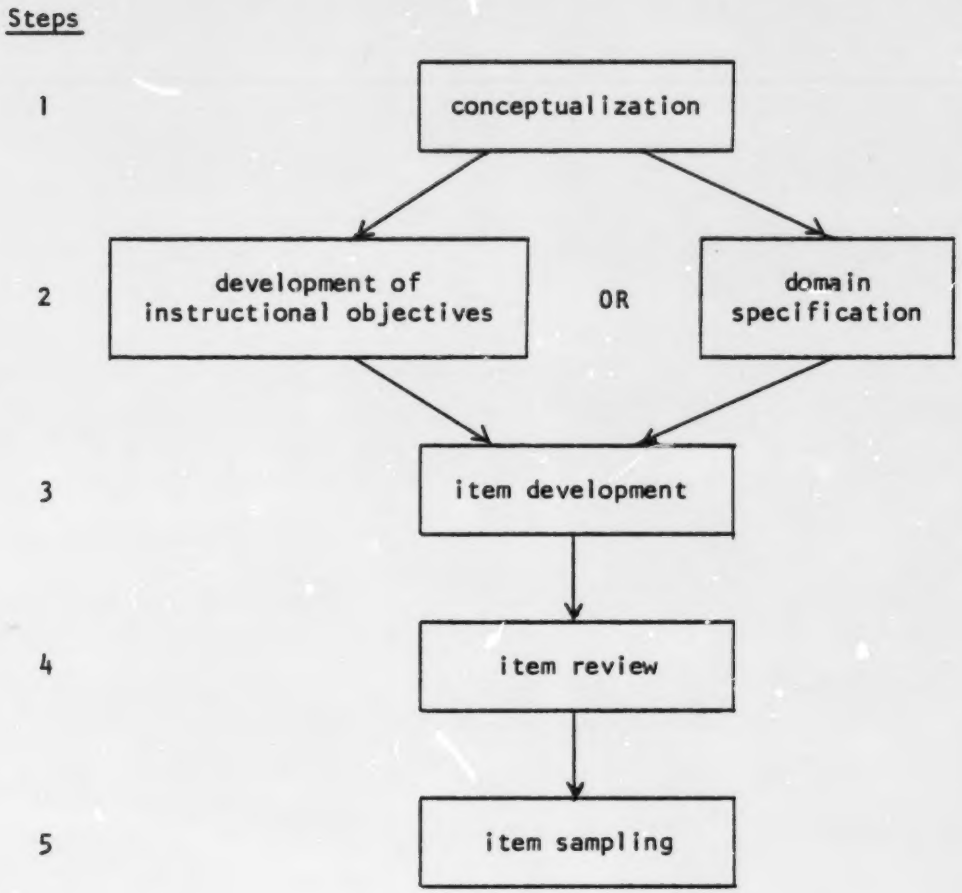

Figure 1. Steps in the development of a CR test.

objectives which represent the behaviors elicited by learners as a result of instruction or (b) the specification of the content domain to be learned. Objectives, for quite some time, have been the mainstay of CR testing, although recent statements like those of Popham's (1975) and Millman's (1974a) have indicated that the inherent weakness in using objectives is that they permit considerable freedom when creating items. Studies like those of Roid and Haladyna (1978a) offer empirical evidence that two item writers may not write items of similar quality if they use only the learning objective as a guide.

In step three, items are developed using one of a number of item writing techniques. The object in step three is to develop a universe or domain of test items which adequately represent the instructional intent as abstractly conceived in step one. A number of methods have been proposed and studied for developing items (e.g., the linguistic-based theory of Bormuth, 1970), and the process is very much in line with the domain specification approach currently advocated by leading test theorists (Hambleton, et al., 1978; Popham, 1975; Millman, 1974a).

While these procedures lead to large numbers of items, Haladyna and Roid (1978) and Hambleton, et al. (1978) have argued for a process whereby items are reviewed either by logical or by empirical means (step four). These item reviews are intended to identify defective items and either revise or discard such items before they are employed in CR testing. Thus, the resultant item domain is one in which logical and empirical reviews have been used to ensure the quality of the domain of items.

The final step in test development (step five) is the selection of items for a CR test. While test blueprints and empirical item selection techniques have been advocated for years, there is strong evidence that random sampling

of items should occur (Millman, 1974b; Popham, 1975; Hambleton, et al., 1978; Haladyna & Roid, 1977). Millman (1974b) provides some guidance on types of sampling plans that may be employed to provide the adequate coverage of the content desired. The practice insures a high degree of content validity.

Within the area of CR testing there are many issues to be studied and resolved. These issues include (a) item review, (b) reliability, (c) decision making, (d) standard setting, and (e) validity. Each of these issues becomes the object of future intensive study and concern. However, the present review is focused on the first three steps of CR test development, which are related to item development. The research and development in the field of CR test-item writing are reviewed, and recommendations are made for future research and development.

Systematic Instruction and Systematic Testing

As test developers are aware, a good CR test is typically used in instruction to monitor student progress with respect to the intent of the instruction. There are a number of instructional systems in the military sector as well as in public education which treat instruction in a scientific manner. That is, instruction is systematic and goal-oriented, and CR testing is used to monitor and certify accomplishment.

As noted earlier, the CR tests that are developed are created by random sampling from a domain of items representing the instruction intent, as illustrated in Figure 1. The important distinction is that two types of CR tests exist--objective-based and domain-based. The objective-based CR test consists of items which were selected or written to reflect a single objective or set of objectives. The domain-based CR test is derived from a specification of the domain and rigid rules for the development of items which do not permit

a great deal of influence by the item writer on the item, thus greatly eliminating the potential for item writer bias. Millman (1974a) has stated that the domain-based CR test is the purest form and a more desirable alternative to the objective-based.

There are several reasons why there has been increasing attention for how items are developed. Foremost among these is that most instructional systems need large collections of CR items in order to provide students with formative and summative evaluation tests that are required. When mastery is not achieved, instructors must provide suitable remediation and retests until mastery is achieved. The consequence of this strategy, which seems to be common to virtually all forms of systematic instruction, is that test items must effectively and logically represent instruction and there must be a great many test items. As Shoemaker (1975) notes, the test item universe and the instruction should be isomorphic.

When item writers create items for CR testing using informal or subjectively inspired methods, or using instructional objectives, they are likely to produce items which vary in quality and difficulty (Bormuth, 1970; Roid & Haladyna, 1978a). The inherent subjectivity in item writing produces a bias that is difficult to overcome in CR testing. Further, defective items are generated, particularly upon the initial draft of a test.

Another reason for concern with item development is that unless test-item writing methods are operationally-defined, these methods cannot be described or documented for other researchers or educators. If the test-item writer uses a mental process that cannot be described and communicated to another educator, the process of item writing remains a private event which is not operationally defined and, hence, not replicable. An operationally-defined method provides a precise description of how items are written so

that two independent item writers using the same method produce virtually identical items. And these items have an integral link with instruction as well as the intent of instruction.

A Classification of Item Writing Methods

There are several major classes of item writing methods that can be contrasted. These range on a continuum from informal-subjective to automated-objective. This state of affairs is illustrated in Figure 2.

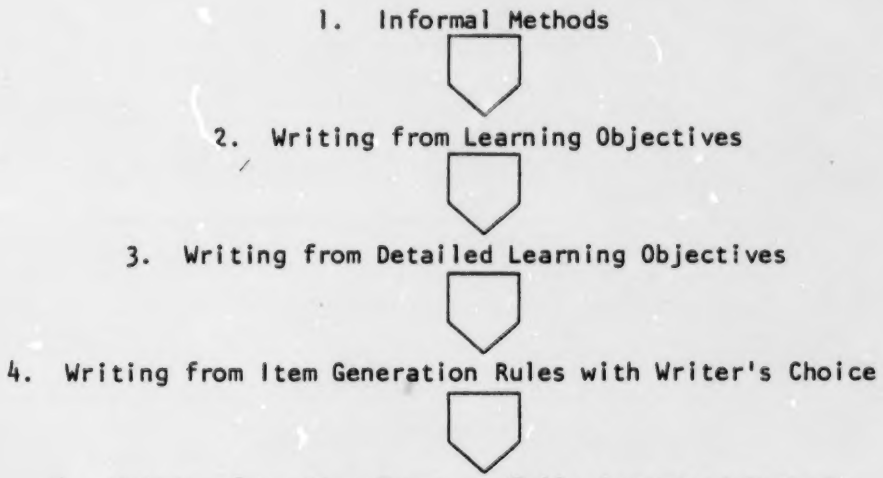
1. Informal Methods
2. Writing from Learning Objectives
3. Writing from Detailed Learning Objectives
4. Writing from Item Generation Rules with Writer's Choice
5. Writing from Item Forms or Fully Automated Methods

Figure 2. A continuum of item writing methods.

The informal methods may involve a listing of the topics to be covered in the content of a course or may simply involve the instructor sitting down and writing items that are felt to be relevant to the course. At levels two and three of the continuum, learning objectives or detailed objectives may be written for a course of instruction, and these are used as guides in

producing test items. At the fourth and fifth levels of the continuum, there may be a domain specification or universe of test items that is defined for a course of instruction and the tests that are used with it (Hively, 1974; Shoemaker, 1975). Since it is assumed that criterion-referenced tests are the appropriate tool for assessing student achievement in systematic instruction, and because these tests are developed using either objectives or domains as the starting point, the emphasis in this review will be on these two classifications of item writing methods.

Objective-Based Methods

Since the appearance of Mager's classic text, Preparing Instructional Objectives (Mager, 1962), there have been a plethora of texts dealing with the subject. The purpose in this review will not be to show how to prepare objectives, but to analyze the role that objectives play in clarifying content and deal with the contribution of objectives to CR test development.

Simply stated: "An objective is an intent communicated by a statement describing a proposed change in a learner--a statement of what the learner is to be like when he has successfully completed a learning experience" (Mager, 1962, p. 3). The key concept in this definition is the "intent" which is the raison dêtre for the objective. Given the objective, the test developer has a good idea what was intended, and s(he) is guided in developing CR test items which are appropriate to this intent. Further, objectives give organization to the content to be learned and are believed to provide focus to learning efforts. In fact, reviews by Hartley and Davies (1976) and by Melton (1978) indicate that the use of objectives does enhance learning, although the latter author warned that the perceived

effectiveness of objectives is an oversimplification in light of the conditions that existed in the research on the effectiveness of objectives.

Studies dealing with item characteristics of CR tests were recently reviewed by Haladyna and Roid (1978). These empirical studies, besides providing a technical base upon which item review may be performed, point to the deficiencies in the approach where objectives are used to generate items. In one study (Roid & Haladyna, 1978a) two item writers used the same learning objectives as a guide in preparing items, but one item writer was found to consistently write more difficult items regardless of the objective. Thus, it seems that the very same subjectivity and bias that is present when the item writer uses his own intuitive notions is present when objectives are used.

It was this dissatisfaction with the anomalies produced by various item writers using objectives that has prompted James Popham, one of the leaders in the use of objectives in teaching and testing, to adopt a modified position which involves domain specifications instead of objectives (Popham, 1975, p. 147). Specifically, a process for clarifying intent through the use of "amplified objectives" was proposed. The object of the amplified objective is to unambiguously provide a domain definition to the item writer. These specifications for items lead item writers to produce items that unambiguously match the intent of instruction.

An example is provided in Figure 3 to illustrate the process whereby objectives can be transformed into amplified objectives and thereby achieve greater clarity of instructional intent.

Instructional Quality Inventory. Another approach to improving objectives is the Instructional Quality Inventory (IQI) developed by the Navy Personnel Research and Development Center, San Diego, and Courseware, Inc.

Descriptive Language: Concrete and Abstract Words Composition Skills

Objective: Given a sentence with a noun or verb omitted, the student will select from two alternatives the word that most specifically or concretely completes the sentence.

Sample Item:

Directions: Mark an "X" through one of the words in parentheses that makes the sentence describe a clearer picture.

Example: The racer (tumbled, went) down the hill.

Amplified Objective:

Stimulus Elements:

1. The student will be given simple sentences with the noun or verb omitted and will be asked to mark an "X" through the one word of a given pair of alternative words that more specifically or concretely completes the sentence.

2. Each test will omit nouns and verbs in approximately equal numbers.

3. Vocabulary will be familiar to a third- or fourth-grade pupil.

Response Alternatives:

1. The student will be given pairs of nouns or pairs of verbs with distinctly varied degrees of descriptive power.

2. In pairs of verbs, one verb will either be a linking verb or an action verb descriptive of general action (e.g., is, goes), and one verb will be an action verb descriptive of the manner of movement involved (e.g., scrambled, skipped).

3. In pairs of nouns, one noun will be abstract or vague (e.g., man, thing), and one noun will be concrete (e.g., carpenter, computer).

Criterion of Correctness:

The correct answer will be "X" marked through the more concrete, specific noun or through the more descriptive action verb in each pair.

Figure 3: Illustrative IOX domain description

(1978). IQI provides a method for examining the consistency between test items, objectives, and instruction. The IQI uses a matrix of test levels by content types that allows the test developer to classify test items and objectives in terms of both task and content. This examination of the relationship between objective and instruction is a positive advance in the technology of item writing, although the degree to which IQI has been successfully implemented is undetermined. Nonetheless, IQI provides a systematic approach to analyzing objectives which may allow for a satisfactory domain specification in the development of CR tests.

Cognitive taxonomies of educational objectives. Three approaches to classifying items and objectives will be briefly presented and reviewed here as a means of clarifying objectives and their CR items. The first one is the most well known, the cognitive taxonomy proposed by Bloom and his colleagues. Others include one proposed by Gagné (1971) and a typology introduced by Williams and Miller (1971).

Bloom's taxonomy (Bloom, Engelhart, Furst, Hill & Krathwohl, 1955) consists of six categories ranging from knowledge, which deals with factual recall, to the highest level, evaluation, which involves judgment. The taxonomy has had tremendous impact on the thinking and practices of educators; any discussion of objectives is imcomplete without reference to this taxonomy. However, seldom are CR tests employed which involve this cognitive taxonomy, and there is little research which recognizes these categories as real domains for objectives and test items. In a recent review of the properties of this taxonomy, Seddon (1978, p. 321) concludes: "No one has been able to demonstrate that these properties do not exist. Conversely, no one has been able to demonstrate that they do." Thus, the

utility of Bloom's taxonomy as a tool for CR test developers still is questionable.

Gagné's approach is largely dependent upon the verb used in the objective. Six categories for objectives are proposed: reinstating, discriminating, identifying, classifying, demonstrating, and generating. The extent to which this approach has been implemented or to which research supports such a classification of objectives is not well known.

A typology was originally introduced by Williams and Miller (1971) which viewed objectives and items particularly as representations of one of five possible types. The term "typology" was used as no order for these categories was implied. A fuller treatment of this work (Miller, Williams & Haladyna, 1978) reveals a system much like Bloom's which, instead, focuses on verbs in test questions as keys to interpreting the cognitive category of behavior in which a test item falls. The categories include factual recall, summarizing, predicting, evaluating, and applying. Williams (1977) added another category, instantiation, which is a derivative of summarizing. His empirical study of the typology revealed that students with a minimum of training could classify test items with a high degree of accuracy. Like other systems for classifying objectives and items, there is little empirical research to support its use, and its usefulness at various educational levels and content areas is unknown.

Item Development Based on Domain Specification

The concept of "domain-referenced" testing was introduced by Hively (1974), although there are earlier descriptions of such an approach to testing (Osburn, 1968). As reported earlier, domain-based tests are derived from content specifications as opposed to objective-based tests which are derived

from instructional objectives. To avoid the semantic confusions that naturally arise from the use of the terms "domain-referenced" and "criterion-referenced," we have chosen to use the more popularized term "criterion-referenced" and treat the general topic in two versions, objective-based CR tests and domain-based CR tests. Clearly, the item development techniques to be reviewed can be subsumed under domain-based CR testing due to the fact that, in this case, rules or algorithms are used to guide the item writer in creating CR test items rather than objectives.

A new technology of domain specification is emerging that provides an alternative to objective-based item writing methods. There are at least five distinctly different approaches to item creation which involve domain specifications. These include: (a) item forms, (b) linguistic-based approaches, (c) facet theory, (d) concept-based testing, and (e) computer-based methods. Each is briefly described and analyzed and research is reviewed which bears on the success with which the method has been employed in CR testing.

Item Forms

Items for a domain-based test may be constructed using a list of rules which specify the structure, format and even some of the wording of the resulting items. These rules are called "item forms" (Hively, 1974) and the total set of items that an item form generates can be considered the "domain" to be measured.

An "item form," as explained by Osburn (1968), "has the following characteristics: (a) it generates items with a fixed syntactical structure; (b) it contains one or more variable elements; and (c) it defines a class of item sentences by specifying the replacement sets for the variable elements." Most of the item forms developed by Osburn and Hively have been

in computational or mathematically-based areas. For example, the following is an item form for a basic mathematics concept:

Item Wording: Which number is second to the smallest of these numbers?

(a), (b), (c), (d), (e)

Elements to Complete the Item: Five numbers, (a) - (e) are provided and the student is required to check the one that is second to the smallest. These numbers must be three-digit integers, with at least two numbers created by rearranging digits in the answer number, and the other numbers with the same "hundreds" digit as the answer number but with randomly generated remaining digits.

Sample Item: Which number is second to the smallest of these numbers?

628, 611, 862, 629, 682

Correct Answer: 628

Millman and Outlaw (1977) have recently applied the concepts of domain-based testing and item forms to tests used in several college courses. Using a small-computer system, they have developed a special programming language that allows an instructor to write an "item program." An item program is a computer program that is capable of producing multiple questions algorithmically, as is done with an item form. The item program specifies a structure for each question. Some of the wording of the item can be fixed and elements in the item can be variables that are replaced to form unique questions. Variable elements can be words or other strings of alphabetic characters, random numbers or quantities computed from mathematical functions. An example of a very simple item form and an item program for an arithmetic problem are shown in Figure 4. The item program is processed by a control program written in BASIC and operates on a small computer. The major

advantage of this system is that only the item program needs to be stored, not the individual items that it generates. A test can be assembled and printed by computer and given to the student at a later time.

Item Form	Item Program
"How much is (X) plus (Y)?"	10 LET X = RANDOM (1, 10)
Where X and Y are integers from 1 to 10.	20 LET Y = RANDOM (1, 10)
	30 QUESTION CONTENT "HOW MUCH IS," X, "PLUS," Y, "?"
	40 ANSWER CONTENT, X + Y

Figure 4. Item form and item program from Millman and Outlaw (1977)

Linguistic-Based Approaches

Bormuth (1970) proposed operationally-defined item writing rules for achievement tests used to assess learning from prose material. These rules are a series of directions which tell an item writer how to rearrange segments of the instruction to obtain the items of that type. Bormuth described two general types of transformations that are possible on prose materials: (a) items derived from sentences and (b) items derived from the relationships between sentences (1970, pp. 39-55). An example of sentence-derived items that test recall of prose material are those produced by the "wh-transformation." Wh-transformation items would be written using a detailed rule summarized as follows: "Select sentences from the instructional materials, replacing a 'wh-pro' word such as who, what, or where for the appropriate part, e.g., subject, in each sentence." For instance, "The test specialist computes the reliability index," could be transformed to: "Who computes the

reliability index?" Wh-items are particularly useful because they can be written to question each separate part of a sentence, and can be made into either multiple-choice or completion format.

Sentence-derived items can also be written to test comprehension of prose material through the use of paraphrasing. Anderson (1972) has emphasized the importance of paraphrasing and has defined it as the case where (a) all substantive words in a sentence are replaced and (b) the original and paraphrased sentences have equivalent meaning.

Items derived from the relations between sentences can be produced, for example, by questioning the cause of an action or result described in prose passages. For instance, the sentences (a) Jim hurt his foot, (b) He was cleaning his gun, and (c) His gun accidentally fired, can be examined for implied causation, resulting in the question, "What caused Jim's hurt foot?" (Bormuth, 1970, p. 54).

Finn (1975) and Roid and Haladyna (1978b) have extended the work of Bormuth by developing question-writing algorithms for prose learning. Finn's original work (1975) involved a rather lengthy and detailed algorithm that contained 82 steps. A streamlined version of the original algorithm was developed by Roid and Finn (1978) and proposed the following important steps in using the algorithm:

1. Analyzing the text;
2. Identifying and classifying key words and their sentences;
3. Transformation of sentences into questions; and
4. Generation of foils for the questions.

Analyzing the text. To develop questions that measure important aspects of a prose passage requires an objective analysis of the text. One approach

to analyzing the prose material for which tests are to be written, is to use a team of experts, teachers or judges to identify the "instructionally relevant" portions of prose instructional materials. Many instructional programs include segments that are directions to the student, explanations of how the program should be used, or other verbal information that is not directly related to the learning objectives for the program. Screening of this material by consensus would be essential for the creation of meaningful, relevant testing.

In addition to screening the passage for instructionally relevant materials, judges could rate each paragraph or sentence as to its relationship to a set of learning objectives that describe the recall and comprehension desired of the reader of the passage. Each sentence could be given numerical codes representing the learning objective to which it is related and the degree of its relevance.

Another approach to text analysis was proposed by Finn (1975), who used a computer analysis of the words in a prose passage. All the words in the passage are keypunched for input to a computer program that does two major tasks: (a) counts the number of times that each word appears in the passage and (b) identifies the standard frequency index of each word. The standard frequency index of each word is a numerical estimate of how often the word occurs in a large corpus of American English (Carroll, Davies & Richman, 1971). The Carroll, Davies and Richman book or its computer-tape version can be used to get the standard frequency index of each word in the passage. The word "the" has the highest standard frequency index of any word, because the average American student is likely to encounter the word "the" once in every ten words of his school book reading. The word "incarnation," for example, has the lowest index because the average student is likely to encounter the

word "incarnation" less often than once in every billion words of his school book reading.

Identifying and classifying key words and their sentences. In order to create test items from prose material in a precise way, individual sentences can be identified and transformed into questions (Bormuth, 1970). One approach to identifying the important sentences in a passage is to have instructors or content experts mark the key sentences in the passage. A consensus of markings can be used to identify the instructionally important sentences. If this consensus is also based on learning objectives for the prose material or instructional program of which it is a part, the method becomes objective-based. The sentence-transformation methods to be described, then make the technique domain-based for two reasons: (a) sentences can be randomly sampled and (b) transformations can be operationally defined. Another approach to identifying key words in prose has been proposed by Finn (1978). The goal of his analysis is to identify "high information" words--words that are relatively rare in American English and occur only a single time in the passage. The sentences in which these high information words occur can then become candidates for transformation into questions which tap important information in the passage. High information words are those that are difficult for students to guess if they are deleted from a prose passage such as is done in a Cloze test (Culhane, 1970). Cloze tests use segments of prose that are presented to a student, usually with every fifth word deleted. The task for the student, then, is to supply the missing words. The easiness with which a word is guessed or filled in by a student on a Cloze test is a measure of the amount of information it provides. The task in a Cloze test is similar to the typical problem in information theory (Shannon & Weaver, 1949) in which a person is receiving a message, but because of noise on the

channel, he is not always sure which message he hears (Finn, 1977). The information in a missing or garbled message is a function of the amount of doubt the receiver has about its identity and is related to the probability of occurrence of certain words or letters. A missing word which is highly predictable, or a common word in the English language, would present less information because students would have less doubt that it completes the prose segment.

Finn (1977) has done considerable research to show that the easiness with which a word is guessed on a Cloze test is predicted by the two important indexes derived from the computer analysis of a passage: (1) the standard frequency index and (2) text frequency. Words that have a low standard frequency index (infrequent in American school books) are typically high in information. However, there is one circumstance in which the information of these words is reduced in relation to a given passage. If the word is repeated frequently (i.e., it has a high text frequency), the information value of that word is diminished and students will supply it more often in a Cloze test following reading of the passage. In other words, repetition of words, even if they are rare in American English, lowers their information value. Candidates for good question words are those which are both rare in American English (have a low standard frequency index) and are singletons in a prose passage (i.e., have a text frequency of 1).

Not all parts of speech are equally good candidates for question words even though they may be high information words. Verbs and adverbs, in particular, are difficult words to remove from a sentence that is transformed into a question. For example, the sentence, "Finn echoed the concern of Bormuth," when transformed to "What did Finn do to the concern of Bormuth?" seems clumsy and seems to be a less important question than the question

"Who echoed the concern of Bormuth?" According to Finn (1978), the most promising parts of speech are adjectives and nouns, or phrases that contain them.

Transformation of sentences into questions. Once an important word and/or its sentence has been identified for a possible question, the sentence must be examined and prepared for transformation. Text and sentences should be analyzed by examining the underlying structure of sentences rather than simply their surface structure. A schema is used to diagram each sentence to identify the important parts of a sentence such as the subject, the verb, the case of the verb, and the object, etc. As Finn has said (1975, p. 344), there are several advantages to basing question-writing procedures on underlying structure analysis rather than on surface structure: (a) expletives, functional verbs, articles and prepositions either appear as part of phrases or they do not appear at all, and, (b) the number of questions possible for a given sentence become independent of the number of words in the sentence and is a function of the number of case phrases and non-zero verbs in the underlying sentence. These are important advantages, because prepositions, articles and many verbs are not desirable as question words in most transformations.

The next step in processing the sentence is to eliminate the question word and to transform the sentence into a question. The question word, usually an adjective, a noun, or a phrase including an adjective or a noun, is removed and is replaced with a wh-word. Where several wordings are possible, an attempt is made to stay as close to the wording of the original sentence as possible. There are a number of further guidelines for preparing sentences for analysis, such as replacing pronouns with their appropriate

nouns or replacing references to previous sentences by clauses or phrases from those references.

The algorithm does not produce 100% agreement among item writers in all cases because of such things as the replacement of phrases from previous sentences. Finn (1975, pp. 357-363) discusses some of the discrepancies among item writers.

Generation of foils for the questions. As is common knowledge among test developers, the writing of good foils for multiple-choice questions is challenging work. The first step in an algorithmic generation of foils is to classify the question words which are the correct answers to the question so that possible foils can be obtained from a list of words in the same classification. The most logical source of foil words would seem to be from the prose passage itself, but in some cases, there may be published lists of words, such as that of Carroll, Davies and Richman (1971), that may be useful sources, if the words cover the same type of subject matter.

The Roid and Haladyna (1978b) study experimented with a foil writing method based on the Carroll, Davies and Richman (1971) list of words. The location of the question word from the prose passage was identified in the Carroll list and then semantically similar words in adjacent parts of the list were identified. These adjacent words, having similar standard frequency indexes, were used as foils. In this particular case, the Carroll list proved to be an unacceptable source, because it produced foils which were too obviously in the wrong subject matter area, and were just plain absurd in many cases. Therefore, Roid and Haladyna developed another method of algorithmic foil construction. The algorithm relies on extracting foils from the prose passage itself. Two variations of the algorithm were developed: one for nouns and one for adjectives.

In the case of nouns, all nouns with a standard frequency index of 60 or less were semantically classified using the method of Frederiksen (1975). For example, some nouns were classified as a concrete animate noun. For a given question word, a random sample of other nouns from the passage that were in the same semantic classification were drawn to create foils.

In the case of adjectives, research on semantic differential technique was used as a basis for constructing a classification of adjectives from the passage (Nunnally, 1967, pp. 536-538). In semantic differential research there are typically three factors that are descriptive of words: (1) evaluation, such as "good," "bad," etc., (2) potency, such as "strong," "weak," etc., and (3) activity, such as "fast," "slow," etc. In addition, Nunnally has defined a fourth factor, "familiarity," such as "simple" or "complex." These four factors were used to group the adjectives in the passage that had standard frequency indexes of less than 60. As a further check on these adjectives, they were subjected to an analysis of whether or not they were "familiar," as determined by the Dale-Chall List of 3,000 Familiar Words (Dale & Chall, 1948). The adjective, to be a candidate for a foil, needed to be absent from that list. The rationale for this was that there are some extremely common adjectives that would be too easy as foils.

Research on these linguistic-based methods for generating items has been mainly limited to a series of studies which trace the development of a method for creating appropriate multiple choice CR test items.

The first of these studies, as described earlier, contrasted objective-based and informal methods of item writing (Roid & Haladyna, 1978a). The objective-based method was more in keeping with the amplified objectives approach. The results of this study showed that while objectives provided

guidelines in the preparation of items, one item writer's items were about 10% more difficult than the other's. The consequence of developing CR test forms based on any particular item writer can be great with respect to misclassifying student examinee performance as adequate or inadequate, as is typically done in many forms of systematic instruction. Thus, item writer bias is a phenomena that affects the difficulty, if not the quality, of CR test items produced informally or with objectives. This study suggests that CR test item writers should not proceed from step one, conceptualization, to item writing or use loosely stated objectives.

In the second of this series of studies, Roid and Haladyna (1978b) examined the effects of variations in linguistic-based algorithms on item characteristics including instructional sensitivity, a criterion measure based on the tendency for items to exhibit change in difficulty as a function of instruction.[1] Four item writers were compared on two methods of selecting sentences, two types of question words, and two foil construction methods. No significant differences between item writers were found on item difficulties, indicating an absence of item-writer bias. Keyword nouns, which are relatively rare words in American textbooks that appear frequently in a prose passage, were found to be unacceptable as question words. Algorithmic methods of foil writing were found to be feasible. Thus, this study indicated that item-writer bias could be eliminated through the use of certain rules dealing with the way sentences are identified and transformed into multiple-choice questions. The study also points to a need for further development in

---

[1] See Haladyna (1974) and Haladyna and Roid (1978) for fuller discussions of this characteristic and measures of it.

the line of the refinement of the algorithms in an effort to achieve more fully automated or even computerized procedures.

In a more recent study in this line of research, Roid, Haladyna and Shaughnessy (1977) examined some of the refinements developed as a result of the previous study in contrast to methods of item writing that are based on paraphrasing of keywords. It was found that passages with greater density (i.e., sections that provided more information) yielded harder items. More importantly, letting item writers have more freedom in the selection of foils produced better items when using the criterion of instructional sensitivity. Verbatim transformations led to items with higher instructional sensitivity. The procedures examined also led to greater control of item difficulty than previously observed. This study documents some of the intriguing advantages of automated multiple-choice item writing methods, and this study also points to the need to further study and improve item generating techniques based on Bormuth's theory of achievement testing. The goal of reducing item-writer bias was effectively achieved, and future work will concentrate at making the process more cost effective and efficient.

Facet Theory[2]

Structural facet theory (Foa, 1958) has existed for some time and has mainly served as a research tool, particularly in the area of attitude measurement. Only recently have there been applications to CR test construction (Engel & Martuza, 1976; Berk, 1978). The purpose of facet theory is to provide the structure and boundaries of a domain of testing conditions. For

[2] The source of the information presented here was mainly derived from excellent presentations of facet theory by Engel and Martuza (1976) and Berk (1978). The reader is referred to these original sources for a more detailed treatment of facet theory for CR tests.

this reason, facet theory is viewed as a method for developing a population of items representing the domain of instruction. The primary advantage of facet theory is that the analysis of content has semantic meaning in a theoretical sense, and there is no need to conduct empirical analyses to search for meaning. The logical analysis of sentences leads to meaningful test items which are easily interpretable. That is not to say that empirical observation is unnecessary in facet theory, but that the theory a priori specifies the nature of the material to be learned and tested. Thus, facet theory, like other similar approaches, allows for an objective specification of the domain which is the target of instruction.

Facet theory specifies the limits of the domain and the orderings of its subparts. In the theory, two aspects are hypothesized: content and statistical. With content, the domain is specified using a semantic structure called the "mapping sentence." This structure is the framework for predicting the statistical structure. The structure is later tested using observations, thus enabling the user to relate theory to observation. In the context of achievement testing, the mapping sentence is a mechanism for defining a content domain and a related set of test items to measure achievement in that domain.

The mapping sentence has fixed and variable parts resembling an item form shell. Parts of the sentence called "facets" are identified which represent some specific information for testing, and the facet elements operate in much the same manner that replacement sets operate in item forms. The sum of all desirable patterns of facets constitutes a facet design.

Using an example provided by Berk (1978), p. 2.):

A. Domain Specification Strategy

1. Item transformation
2. Item forms
3. Algorithms
4. Mapping Sentences

} is most appropriate for defining

B. Content Domain

1. reading
2. language
3. mathematics
4. science
5. social studies

The sentence has facets of domain specification and content domain. The elements of each facet are ordered in some meaningful way. One refers to rules for choosing facets and their elements (McGrath, 1967). These rules are summarized from Berk's paper (1978):

1. Objects should be classified by all properties or facets.

2. Each facet should be divided into an exhaustive set of categories or elements.

3. The elements should be mutually exclusive; that is, each element is classifiable into one and only one category.

4. The logical relationship among elements of a facet should be specifiable.

5. The logical relationship among facets should be specified.

6. The facets should exhaust the domain of interest.

From the example given above, there are a total of twenty combinations of elements from Facet A and Facet B. Thus, twenty statements exhaust the

domain of possibilities, and twenty true-false items are possible. For example: Item transformation is most appropriate for defining mathematics. This is A1B3, which is false.

This particular facet design is useful for true-false or completion formats. Building a multiple-choice domain of items is considerably more complex. Besides identifying the correct answer for a particular facet, the additional burden is to produce three or four plausible foils. This process, described by Berk (1978), involves a logical analysis of potential distractors which are drawn from the elements of the facet. Using the previous example with some modification:

Item transformations are most appropriate for defining:

a. social studies

b. language

c. reading

d. mathematics

The benefits of facet designs, as well as other similar approaches were discussed by Engel and Martuza (1976):

1. Both item stem and foils are systematically constructed.

2. Facet design is based on a theory of content and how content is defined. Therefore, the identification of foils occurs in the context of how foils are more or less attractive as incorrect responses. As a consequence, incorrect responses have meaningful interpretations in a diagnostic vein.

3. The procedures provide a logical connection between content and the multiple-choice item.

4. Items produced may be logically compared with respect to content difficulty and appropriateness, thus making the construction of parallel

test forms easier and less subject to capriciousness which exists when random sampling is used to create items.

As noted earlier, facet theory is a relatively young field of content specification for domain-based CR tests. There is very little known about its applicability to various subject matters or the empirical characteristics of tests constructed using facet designs.

Engel and Martuza (1976) conducted an empirical study of facet designs. The procedures led to functionally equivalent parallel forms tests. Further, results indicated that the method works equally well with highly structured material like mathematics as well as more abstract material. Finally, the study showed the feasibility of facet theory as a method for domain-based CR test item construction. It is also interesting to note that like amplified objectives, the facet design can be used with an objective as the mapping sentence, thereby capitalizing on existing objectives.

Like other approaches to domain-based testing, facet theory appears quite promising. These seminal works by Engel and Martuza (1976) and Berk (1978) provide a clear picture of the nature and potential of facet theory. There remains, however, much work to be done to expand the theory and to apply it to various subject matters as well as to compare it to other domain-based approaches in an effort to uncover which set of procedures is most efficient, feasible, and defensible in light of the abstract conceptualization of instructional intent posited as the first step in CR test development. As with other approaches, more development coupled with empirical research should reveal the utility of facet theory and its eventual role in CR testing.

Concept-Based Testing

The work of Markle and Tiemann on the teaching and testing of concepts is an excellent example of domain-based testing that goes beyond the factual level of learning (Markle & Tiemann, 1974; Markle, 1975; and Tiemann, Kroeker, & Markle, 1977). They have defined concepts as classes of objects, events, or relations which vary among themselves and yet are all grouped together and called by the same name. A student's understanding of a concept is tested by checking for generalization to new examples and discrimination of non-examples. Examples and non-examples are used in teaching a concept and then a new set of examples and non-examples are used to test the student's understanding of the concept. If we were teaching the concept "chair," we might use the examples of a wooden chair and an upholstered chair and the non-examples of a stool and a love-seat in the teaching exercise. In testing for understanding of the concept of chair we might use the examples of a rocking chair and a rattan chair, and the non-examples of a bench and a chaise longue.

Markle and Tiemann (1969) provide guidelines and many practical examples of the analysis of concepts. The analysis of concepts involves listing the irrelevant and critical attributes of the concept. An irrelevant attribute is a property of any particular example which can be varied without changing an example to a non-example. For instance, the number of legs is irrelevant to the concept chair, because we can have a modernistic chair with a pedestal or a standard four-legged chair. Critical attributes are properties of every example of the concept, and if they are removed, the example becomes a non-example. For instance, possession of a back, a rigid seat, and the requirement of a "single-person seat" are the critical attributes of chair. Irrelevant

attributes of chair are whether it has rockers, arms, the size of the back, etc. After the critical attributes and irrelevant attributes have been listed, and lists of examples and non-examples have been generated, it then becomes possible to construct domain-based criterion-referenced tests for a given concept, a series of concepts, or a principle (a statement of relationship between concepts). Such a test would be constructed by choosing a random sample of examples and non-examples and systematically varying critical attributes and irrelevant attributes. An example of a concept analysis and a sample item for the concept "antonym" is given in Figure 4.

Markle and Tiemann recently extended their work to the case of multiple coordinate concepts (Tiemann, Kroeker & Markle, 1977). An example of coordinate concepts are the four contingencies of positive and negative reinforcement and positive and negative punishment. In this case, the four concepts overlap to the point where an example of one concept is a non-example of the other concept. Students need to learn to reject a non-example in one case but accept it as an example of another concept that is related. The Tiemann, Kroeker and Markle study provides an example of how a domain-based test is produced for the set of four coordinate concepts by systematically sampling examples of each of the concepts, varying them on their attribute dimensions.

Computer-Based Methods

Computers have been used as aids in assembling or administering tests for many years (e.g., Atkinson & Wilson, 1969). Many attempts centered on the use of item banks containing all of the actual items from which samples were drawn for testing. More sophisticated systems included the capability of items being composed such as was done in the mid-1960's in the drill-and-practice exercises of the Stanford computer-assisted instruction project

I. Grammar Concept: Antonym

A. Attributes

	Critical	Irrelevant
1. along some continuum or in some dichotomy, has a meaning opposite to the meaning of some other (given) word.	X	
2. is the same part of speech as the first word.	X	
3. is a new word, not a variation of the first word.	X	
4. has the same function in a sentence as the first word.	X	
5. may be drawn from various parts of speech--nouns, verbs, pronouns, adverbs, adjectives, or prepositions.		X
6. equivalence in length.		X

B. Examples

Teaching Examples		Teaching Non-Examples	
1. bad; good	(5, adjective)	1. vain; proud	lacks 1
2. danger; safety	(5, noun)	2. reason; motive	lacks 1
3. live; die	(5, verb)	3. we; us	lacks 1
4. he; she	(5, pronoun)	4. above; upon	lacks 1
5. rapidly; slowly	(5, adverb, 6)	5. merrily; sad	lacks 2 & 4
6. in; out	(5, preposition)	6. happy; unhappy	lacks 3
		7. capable; incapable	lacks 3
		8. disputable; agree	lacks 4 & 2

Testing Examples		Testing Non-Examples	
1. hot; cold	(5, adjective)	1. imaginary; fanciful	lacks 1
2. loss; gain	(5, noun)	2. chair; couch	lacks 1
3. elevate; lower	(5, verb, 6)	3. behind; next to	lacks 1
4. you; me	(5, pronoun)	4. gloom; bright	lacks 2 & 4
5. gaily; sadly	(5, adverb)	5. violent; non-violent	lacks 3
6. over; under	(5, preposition)	6. valid; invalid	lacks 3
		7. weak; forcibly	lacks 4 & 2

C. Sample Test Item

Which of the following pairs of words are antonyms?

a. imaginary -- fanciful
*b. elevate -- lower
c. valid -- invalid
d. weak -- forcibly

Correct Answer: b

Figure 4: Example of a concept analysis used to develop domain-referenced tests of concept learning.

(Suppes, Jerman & Groen, 1966). Systems with the capability of generating items can be used to implement domain-based testing, and, for that reason, they will be described in more detail in the section that follows.

Most of the major computer-assisted instruction author languages have the capability of producing algorithms for domain-based test items. Several of the CAI languages discussed by Roid (1974) such as COURSEWRITER, PLANIT and TUTOR have functions which allow an item program to be written as described by Millman and Outlaw (1977). For example, Atkinson (Atkinson & Wilson, p. 153) used COURSEWRITER to program reading exercises and criterion tests for the Stanford initial reading programs. A sample exercise is the sentence, "Dan saw the ____ hat," for which the student is to choose one of a set of computer-assembled words such as "tan," "fat," "man" or "run." The fill-in answers are selected by rules from words previously presented in lessons to the student.

Another example is the work of Fremer and Anastasio (1969) who developed computer-generated items for testing spelling. They conducted an analysis of types of misspellings which are used by writers of spelling items. A set of error generation rules were developed and programmed for a computer using the SNOBOL programming language. SNOBOL is a language developed at Bell Telephone Laboratories especially to handle strings of symbols or alphabetic characters. Error generation rules included the inversion of letters within a word, omission of letters, substitution or insertion of letters. An example for the word "preferable" would be "perferable" or "preforable" or "preferabal." Fremer and Anastasio found that computer-generated lists of spelling items were judged highly useful by a panel of spelling test developers.

Beginning with the pioneering work of Hively, Patterson and Page (1968), a great deal of work has been done on domain-based tests in mathematics. For example, Hsu and Carlson (1973) developed item-generation routines used to help construct tests for the elementary mathematics level of the Individually Prescribed Instruction program at the University of Pittsburgh. They used the concept of item forms developed by Hively in programming item-generation routines in FORTRAN IV for a PDP-10 time-sharing computer system. Hsu and Carlson make the important suggestion that the difficulty level and discrimination of item forms be tested by collecting data from tryouts of each item form. Because individual test items are automatically produced, the only way to insure the quality of test items is to improve the quality of the item forms. By field testing and keeping item statistics at the item form level, it will be possible to develop higher quality domain-based tests.

Beginning with the work of Osburn (1968), a number of university professors have developed computer-generated testing systems, particularly in the sciences. For example, Johnson (1973) has developed a system for computer-generated repeatable examinations in chemistry at the college level. His system includes control programs and subroutines written in FORTRAN IV. Each subroutine defines an item form. These item forms include numerical constants which are randomly generated by computer or variable wordings which might include different names of chemical compounds which are substituted into various versions of each individual item. An example of an item form and an individual item from Johnson's system is given in Figure 5.

A number of technological developments have recently improved the efficiency of computer-aided testing systems. For instance, McClain, Wessels and Sando (1975) described the unique advantages of Iowa's "Item

Item Form

IF _____ ML. OF _____ MOLAR _____ IS MIXED WITH _____ ML. OF _____ MOLAR _____, THE FINAL SOLUTION WILL BE:

A) _____ MOLAR IN _____

B) _____ MOLAR IN _____

C) _____ MOLAR IN _____

D) _____ MOLAR IN _____

E) _____ MOLAR IN _____

Sample Item

IF 43.6 MLS. OF 1.50 MOLAR NAOH IS MIXED WITH 38.5 MLS. OF 1.14 MOLAR HNO3, THE FINAL SOLUTION WILL BE:

A) 1.33 MOLAR IN OH (-)

B) 1.10 MOLAR IN H (+)

C) 0.260 MOLAR IN H (+)

D) 1.33 MOLAR IN H (+)

E) 0.260 MOLAR IN OH (-)

Figure 5. Example of an item form and item from Johnson's Computer-Generated Repeatable Chemistry Exam System (Johnson, 1973).

Pool System for Instructional Management." This system was written in PL/1 (Programming Language 1) for the IBM 360 and includes an item-generation capability. The prototype of an item is stored in a pool from which the system can retrieve it on demand. The prototype states the rules for creating the specific variable components in both the question and its multiple-choice foils. As a somewhat unique part of this system, the functions to be used in generating the numerical values or variable elements in the item form are coded in with the prototype item. When the prototype is retrieved from the pool, the computer program then processes the prototype and its imbedded rules in order to generate a specific item or items. This system has the unique advantage of having the item-generating algorithm somewhat independent from the control program. That is, instead of having specific subroutines for each algorithm, a coding system is developed for placing the algorithmic rules into the item form in a coded fashion. Then the control program simply needs to interpret those codes in order to implement the algorithm. This would seem to lead to less "overhead" in computer programming.

Other computer-based item generation programs have been described by Olympia (1975) and Vickers (1973). The work of Vickers is interesting in that it involves the computer generation of questions useful in the teaching of FORTRAN programming. A series of subroutines, that employ random number generators, are used to compose FORTRAN-like statements or variables. Then, the student might be asked to discriminate between correct and incorrect statements or to classify types of variables. This is an excellent example of a sophisticated item writing method that produces a virtually unlimited number of items.

In summary, there is a wide variety of examples where computerized item writing methods have been implemented. Many of these methods are in use in college and university courses, particularly in the sciences. The capability to implement these methods is available on most of the major computer installations in the nation. Thus, the technology is available for creating domain-based criterion-referenced tests. The challenge that remains is in the definition of domains and the creation of item-writing algorithms in a wide variety of subject matter areas. Also, creativity is required to develop domains at the conceptual as well as factual levels of learning.

Conclusions and Recommendations

Instructional development has been capably served by several principles of learning and testing which involve the use of instructional objectives. Research on systematic instruction (reviewed by Block, 1971; Hartley & Davies, 1976; Robin, 1976 among others) has been impressive. Both systematic instruction and the use of objectives appear to have positive effects on learning. Unfortunately, objectives permit much too much freedom to the often inexpert item writer which in turn results in many items which are instructionally irrelevant or psychometrically unsound. Despite logical and empirical methods of item review, many of the problems in producing items may be avoided by employing one of several domain-based item generating approaches. The five reviewed here are fairly sound in theory, research and development. Preliminary findings indicate a vast potential for the creation of large groups of items which may form the basis of sound CR testing in the future.

The key to all this activity is the acceptance of the process in CR test development illustrated in Figure 1. Test theorists and practitioners are in accord when they maintain a concern for a logical and close correspondence between instruction and testing. Because domain-based methods elaborate on objective-based methods, it is possible to achieve almost perfect objectivity in the creation of test items. Therefore, the item writing methods being reviewed promise a more scientific approach to item development, which in turn improves the measuring of student achievement in systematic forms of instruction.

References

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Atkinson, R. C., & Wilson, H. A. Computer-assisted instruction: A book of readings. New York: Academic Press, 1969.

Berk, R. A. An application of structural facet theory to objective based achievement testing. Paper presented at the annual meeting of the Eastern Educational Research Association, Williamsburg, Virginia, March 1978.

Block, J. H. (Ed.) Mastery learning: Theory and practice. New York: Holt, Rinehart, and Winston, Inc., 1971.

Bloom, B. S. (Ed.), Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. Taxonomy of educational objectives. New York: Longman's, Green & Company, Inc., 1956.

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Carroll, J. B., Davies, P., & Richman, B. Word frequency book. Boston: Houghton-Mifflin, 1971.

Culhane, J. W. CLOZE procedures and comprehension. The Reading Teacher, 1970, 23, 410-413.

Dale, E., & Chall, J. W. A formula for predicting readability. Educational Research Bulletin, 1948, 27, 11-28.

Engel, J. D., & Martuza, V. R. A systematic approach to the construction of domain-referenced multiple-choice test items. Paper presented at the meetings of the American Psychological Association, Washington, D.C., September 1976.

Finn, P. J. A question writing algorithm. Journal of Reading Behavior, 1975, 4, 341-367.

Finn, P. J. Word frequency, information theory and cloze performance: A lexical-marker, transfer-feature theory of processing in reading. Unpublished paper, State University of New York at Buffalo, School of Education, 1977.

Finn, P. J. Generating domain-referenced, multiple-choice test items from prose passages. Paper presented at the meetings of the American Educational Research Association, Toronto, March 1978.

Foa, U. G. New developments in facet design and analysis. Psychological Review, 1965, 72, 272-274.

Frederiksen, C. H. Representing logical and semantic structure of knowledge acquired from discourse. Cognitive Psychology, 1975, 7, 371-458.

Fremer, J., & Anastasio, E. J. Computer-assisted item writing -- I (spelling items). Journal of Educational Measurement, 1969, 6, 69-74.

Gagné, R. M. Defining objectives for six types of learning. (Washington, D.C.: American Educational Research Association, 1971.) Adapted with the author's permission from a research training tape prepared for AERA.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 93-99.

Haladyna, T., & Roid, G. R. An empirical comparison of three approaches to achievement testing. Paper presented at the annual meeting of the American Psychological Association, San Francisco, 1977.

Haladyna, T., & Roid, G. The role of instructional sensitivity in the empirical review of criterion-referenced tests. Monmouth, OR: Teaching Research, 1978.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Hartley, J., & Davies, I. K. Preinstructional strategies: The role of pretests, behavioral objectives, overviews and advance organizers. Review of Educational Research, 1976, 46, 239-265.

Hively, W. Introduction to domain-referenced testing. Educational Technology, 1974, 14, 5-10.

Hively, W., Patterson, H. L., & Page, S. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Hsu, T., & Carlson, M. Test constructional aspects of the computer assisted testing model. Educational Technology, 1973, 13(3), 26-27.

Instructional Quality Inventory: Interim Training Manual. San Diego, CA.: Navy Personnel Research and Development Center, 1978.

Johnson, K. J. Pitt's computer-generated chemistry exam. In Proceedings of the Conference on Computers in Undergraduate Curricula, 1973, 199-204.

Mager, R. F. Preparing instructional objectives. Palo Alto, California: Fearon Publishers, 1962.

Markle, S. M. They teach concepts don't they? Educational Researcher, 1975, 4(6), 3-9.

Markle, S. M., & Tiemann, P. W. Really understanding concepts. Champaign, Illinois: Stipes Publishing Company, 1969.

Markle, S. M., & Tiemann, P. W. Some principles of instructional design at higher cognitive levels. In R. Ulrich, T. Stachnik & T. Mabry (Eds.), Control of human behavior (Vol. III). Glenview, Illinois: Scott-Foresman, 1974, pp. 312-323.

McClain, D. H., Wessels, S. W., & Sando, K. M. IPSIM--Additional system enhancements utilized in a chemistry application. Proceedings of the Conference on Computers in Undergraduate Curricula, 1975, 139-145.

McGrath, J. E. A multi-facet approach to classification of individual, group, and organizational concepts. In B. Indik and K. Berrien (Eds.), People, groups, and organizations: An effective integration. New York: Teachers College Press, 1967, pp. 191-215.

Melton, R. F. Resolution of conflicting claims concerning the effect of behavioral objectives on student learning. Review of Educational Research, 1978, 48, 291-302.

Miller, H. G., Williams, R. G., & Haladyna, T. M. Beyond facts: Objective ways to measure thinking. Englewood Cliffs, N.J.: Educational Technology, 1978.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, California: McCutchan Publishing Company, 1974. (a)

Millman, J. Sampling plans for domain-referenced tests. Educational Technology, 1974, 14, 17-21. (b)

Millman, J., & Outlaw, W. S. Testing by computer. Ithaca, New York: Cornell University Extension Publications, 1977.

Nunnally, J. Psychometric theory. New York: McGraw-Hill, 1967.

Olympia, P. L., Jr. Computer generation of truly repeatable examinations. Educational Technology, 1975, 14(6), 53-55.

Osburn, H. G. Item sampling for achievement testing. Educational and Psychological Measurement, 1968, 28, 95-104.

Popham, W. J. Educational evaluation. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Robin, A. Behavioral instruction in the college classroom. Review of Educational Research, 1976, 46, 313-354.

Roid, G. H. Selecting CAI author languages to solve instructional problems. Educational Technology, 1974, 14(5), 29-31.

Roid, G., & Finn, P. Algorithms for developing test questions from sentences in instructional materials. San Diego, California: Navy Personnel Research and Development Center, 1978.

Roid, G. H., & Haladyna, T. M. A comparison of objective-based and modified-Bormuth item writing techniques. Educational and Psychological Measurement, 1978, 35, 19-28. (a)

Roid, G., & Haladyna, T. A comparison of several linguistic-based, multiple-choice item writing algorithms. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978. (b)

Roid, G., Haladyna, T., & Shaughnessy, J. Item writing for domain-referenced tests of prose learning. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Seddon, G. M. The properties of Bloom's taxonomy of educational objectives for the cognitive domain. Review of Educational Research, 1978, 48, 303-323.

Shannon, C. E., & Weaver, W. The mathematical theory of communication. Urbana: University of Illinois Press, 1949.

Choemaker, D. M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45, 127-148.

Suppes, P., Jerman, M., & Groen, G. J. Arithmetic drills and review on a computer-based teletype. Arithmetic Teacher, 1966, April, 303-308.

Tiemann, P., Kroeker, L. P., & Markle, S. M. Teaching verbally-mediated coordinate concepts in an on-going college course. Paper presented at the meetings of the American Educational Research Association, New York, April 1977.

Vickers, F. D. Creative test generators. Educational Technology, 1973, 13(3), 43-44.

Williams, R. G. A behavioral typology of educational objectives for the cognitive domain. Educational Technology, 1977, 17(6), 39-46.

Williams, R. G., & Miller, H. G. Constructing higher level multiple-choice questions covering factual content. Educational Technology, 1973, 13(5), 39-42.